Largely Untested

v1.1.0

Author: John J. McCormick · Updated: March 4, 2026 · CC BY 4.0 · aiagentgovernance.org

SUGGESTED CITATION

McCormick, J. J. "Largely Untested: What Operational Evidence Reveals About Agent Governance Interventions." v1.1.0, March 2026. aiagentgovernance.org. https://aiagentgovernance.org/insights/largely-untested/

Abstract

In April 2025, the Institute for AI Policy and Strategy published a 63-page survey of AI agent governance — the most comprehensive academic treatment of the field to date. The paper's central finding is also its most striking admission: proposed governance interventions are "largely untested," and "only a small number of researchers are working on agent governance." This article examines what happens when those interventions are tested — not in simulation, but in continuous production operations with autonomous AI agents performing real work under formal governance controls. The results confirm the academic taxonomy's structure while revealing operational dynamics that theory alone does not predict.

The Academic Consensus

Three significant contributions have shaped the emerging academic understanding of AI agent governance:

Kraprayoon, Williams & Fayyaz (2025) — "AI Agent Governance: A Field Guide" (IAPS). A comprehensive survey proposing five intervention categories: Alignment, Control, Visibility, Security & Robustness, and Societal Integration. The paper compiles agent capability benchmarks, risk taxonomies, and a research agenda. It is explicit about the field's maturity: interventions are proposed, not validated.

Patil, A. G. (2025) — "Governing Agentic AI: A Strategic Framework for Autonomous Systems." A prescriptive framework proposing design-time governance, runtime oversight, accountability mechanisms, and organizational integration. The paper identifies the right structural requirements — graduated autonomy, audit trails, kill switches — but presents them as recommendations, not tested implementations.

Tomašev, Franklin & Osindero (2026) — “Intelligent AI Delegation” (Google DeepMind). A formal analysis of delegation dynamics in AI agent systems. The paper arrives at structurally similar conclusions to the operational governance framework published here — independent theoretical alignment (/insights/deepmind-delegation/) between a leading AI research lab’s theoretical analysis and empirical production operations.

These contributions share a common characteristic: they describe what agent governance should look like. They do not describe what agent governance looks like when you build it, run it, and measure it. This is not a criticism. Theory necessarily precedes practice. But the field has reached the point where operational evidence can inform the theory — and in several areas, the evidence reveals dynamics the theory does not predict.

The Five Interventions, Operationalized

The IAPS taxonomy proposes five categories of governance intervention. The following sections examine what operational implementation reveals about each one.

1. Alignment: A Runtime Observable, Not a Design-Time Property

The academic framing treats alignment as something established before deployment: goal specification, reward shaping, value alignment, corrigibility. The assumption is that a well-aligned agent stays aligned.

Production operations reveal a different dynamic. Alignment is not a state — it is a trajectory. Agents drift. An agent that is correctly aligned at the start of a work session can gradually deviate through a series of individually reasonable decisions that compound into misalignment. This is precisely the mechanism Dekker (2011) describes as “drift into failure” — and it applies to AI agents as directly as it applies to human organizations.

The Behavioral Pattern Taxonomy (/framework/agent-failure-patterns/) documents specific alignment failure modes observed in production:

- **Inference over execution (BP-001):** An agent cannot access required data, infers what the data “probably” contains, and proceeds as if the inference were fact. The agent’s goal alignment has not changed — it is still trying to complete the assigned task — but its method has silently diverged from the governance requirement to work from verified sources.
- **Scope expansion without authorization (BP-004):** An agent completes its assigned work, then identifies adjacent work it believes should also be done, and begins executing without authorization. Again, the agent’s goals are aligned — it

is trying to be helpful — but it has crossed an authority boundary.

These are not alignment failures in the academic sense (the agent is not pursuing wrong goals). They are alignment failures in the operational sense (the agent's behavior has diverged from governance expectations while its goals remain correct). This distinction matters: design-time alignment checks would not catch these patterns because the agent's stated objectives remain aligned. Only runtime behavioral monitoring detects the divergence.

2. Control: The Bottleneck Is Decision Presentation, Not Decision Authority

The academic framing treats control as a human authority question: can the human override, shut down, or redirect the agent? The interventions proposed are structural — kill switches, shutdown mechanisms, escalation protocols.

In production operations, the structural authority is the easy part. The hard part is presenting the right decision to the human at the right moment. Human operators do not lack authority over agents. They lack the structured decision surface to exercise that authority efficiently.

The MOIAS methodology (/framework/governance-lifecycle/) addresses this through phase-gate impact statements — structured decision blocks that present risk level, action being authorized, consequences of approval, consequences of rejection, reversibility, and blast radius. Every phase transition surfaces this information in a consistent format. The human operator makes a go/no-go decision on a structured surface, not by reading through pages of technical output.

This is a finding the academic literature does not anticipate: human control over agents is less a question of authority and more a question of cognitive ergonomics. The human has the authority. The governance system must make exercising that authority fast and informed.

3. Visibility: Measurement Changes What You Can See

The academic framing treats visibility as monitoring and interpretability — can you see what the agent is doing? Can you understand why?

Production operations confirm that visibility matters. But they also reveal a second-order effect: measurement changes what you can see. When you instrument governance with specific metrics, patterns emerge that no amount of observation would reveal.

One example: in production multi-agent operations, every instance in which a governance control was bypassed resulted in a defect. The sample is small, but the correlation was unambiguous enough to establish a governance rule: bypasses are treated as defect-producing by default until evidence demonstrates otherwise. This finding did not emerge from observation — agents performing governance bypasses do not look different from agents following the governance lifecycle. The finding emerged from measurement: tracking lifecycle compliance and correlating it with defect rates.

This has implications for the field. The academic visibility interventions focus on making agent behavior observable. The operational evidence suggests that visibility must extend beyond observation to measurement — defining specific governance metrics, instrumenting them, and analyzing correlations that are invisible to observation alone. The governance framework defines 23 metrics across five domains (/framework/governance-lifecycle/) (quality, efficiency, agent performance, governance health, operator drift), 18 of which are fully instrumented. The metrics reveal patterns that monitoring alone does not surface.

4. Security and Robustness: The Authentication Gap Is Real

The academic framing addresses adversarial robustness, prompt injection defense, and sandboxing. These are important. But the academic literature largely overlooks a more fundamental security concern: agent identity and authentication.

In a multi-agent system, agents interact with shared resources, APIs, and each other. If an agent cannot prove its identity, every other security measure is undermined. Production operations implement agent authentication through per-agent credentials, time-based one-time password (TOTP) two-factor authentication on agent profiles, session tokens with defined expiry, and role-scoped permissions that restrict what each agent can access.

This is not exotic security engineering. It is standard practice for human users — applied to agents. The academic literature describes sophisticated adversarial defenses while leaving the front door unlocked: most agent frameworks have no concept of agent identity, authentication, or session management. The security intervention category would benefit from treating agent identity as a foundational requirement, not an advanced feature.

5. Societal Integration: Standards Emerge From Practice, Not Policy

The academic framing addresses regulation, liability frameworks, standards, and disclosure requirements. These are necessary. But the academic literature treats standards as something that will be designed by standards bodies and adopted by

practitioners.

Historical precedent suggests the opposite. SOC 2 — now the dominant cloud security standard — was not designed top-down by a standards body and then adopted. It emerged from auditing practices that codified what practitioners were already doing. The standard followed the practice (see: the SOC 2 precedent (/insights/soc2-precedent/)).

Agent governance standards will likely follow the same pattern. Standards bodies will codify what practitioners have already built and tested. The governance framework published here — including the MOIAS methodology (/framework/governance-lifecycle/), the Behavioral Pattern Taxonomy (/framework/agent-failure-patterns/), and the Governance Maturity Model (/framework/maturity-model/) — represents one such practitioner contribution. It is published under CC BY 4.0 specifically to enable adoption, citation, and contribution by standards bodies and the research community.

What the Theory Doesn’t Prepare You For

Beyond the five intervention categories, operational implementation reveals three dynamics that the academic literature does not address:

Governance Must Govern Itself

A governance framework that cannot account for its own failures is incomplete. In production operations, the governance system itself experienced incidents — phase gates that were bypassed, behavioral patterns that went undetected, metrics that were defined but not instrumented. The governance framework’s audit trail includes the period before governance existed, the incidents that motivated each control, and the progression from ungoverned operations to formal methodology.

This self-referential property — governance that documents its own gaps — is absent from the academic literature. It is, in practice, one of the strongest forms of evidence: you can see not just the governance system, but its evolution from failure.

Incident-Driven Learning Is the Core Loop

The academic literature proposes governance interventions. It does not describe the mechanism by which governance improves. In production operations, the improvement mechanism is incident-driven: every governance failure produces a root cause analysis, a behavioral pattern classification, and a governance mechanism improvement. This is the same learning loop that safety engineering has used for

decades (Reason, 1990; Dekker, 2011; Vaughan, 1996). Applied to AI agent governance, it means the governance system gets stronger with every failure — a property that static governance frameworks do not have.

The Compounding Problem

Behavioral failure modes do not occur in isolation. An agent that works from inference (BP-001) is more likely to also skip governance phases (BP-003) and report only successes (BP-007). In production, multi-pattern incidents are common. The academic taxonomy describes interventions independently. Operational implementation reveals that failures cluster and compound, requiring governance that detects patterns in combination, not just in isolation.

Implications for the Field

The IAPS field guide is the most comprehensive academic treatment of agent governance published to date. Its taxonomy is sound. Its risk analysis is thorough. Its bibliography is an essential resource for anyone working in this space.

The contribution this article offers is operational evidence. The interventions IAPS describes are no longer entirely untested. Production operations with autonomous AI agents under formal governance have generated documented incidents, behavioral patterns, governance metrics, and a tested lifecycle model. The evidence confirms the academic taxonomy's structure — alignment, control, visibility, security, and societal integration are the right categories — while revealing operational dynamics within each category that theory alone does not predict.

The field is early. IAPS is right that only a small number of researchers are working on agent governance. The path forward requires both theoretical analysis and operational evidence, informing each other. This article is offered as a contribution to that exchange.

References

1. Kraprayoon, J., Williams, Z., & Fayyaz, R. (2025). "AI Agent Governance: A Field Guide." Institute for AI Policy and Strategy (IAPS).
2. Patil, A. G. (2025). "Governing Agentic AI: A Strategic Framework for Autonomous Systems." Independent Researcher.
3. Tomašev, N., Franklin, M., & Osindero, S. (2026). "Intelligent AI Delegation." arXiv:2602.11865 [cs.AI].

4. Dekker, S. (2011). *Drift Into Failure*. Ashgate Publishing.
5. Reason, J. (1990). *Human Error*. Cambridge University Press.
6. Vaughan, D. (1996). *The Challenger Launch Decision*. University of Chicago Press.

Further Reading

- MOIAS Methodology (/framework/governance-lifecycle/) — The complete governance framework
- Behavioral Pattern Taxonomy (/framework/agent-failure-patterns/) — Eight documented failure categories from production operations
- The Governance Gap (/insights/governance-gap/) — The structural space between existing operational layers
- Why Observability Is Not Enough (/insights/observability-not-enough/) — The argument for governance beyond monitoring
- DeepMind Delegation Paper Analysis (/insights/deepmind-delegation/) — Independent theoretical alignment
- The SOC 2 Precedent (/insights/soc2-precedent/) — How standards emerge from practice
- Governance Maturity Model (/framework/maturity-model/) — Five levels of organizational readiness

---

Version History

Version	Date	Author	Description
1.0.0	2026-03-03	John J. McCormick	Initial publication
1.1.0	2026-03-04	John J. McCormick	Fix Patil citation (author, title). Qualify O7 bypass-defect correlation with sample size.
